

Spk2SRIImgNet: Super-Resolve Dynamic Scene from Spike Stream via Motion Aligned Collaborative Filtering Supplementary Material

Yuanlin Wang^{1,2}, Yiyang Zhang^{1,2}, Ruiqin Xiong^{1,2,*}, Jing Zhao^{1,2},
Jian Zhang³, Xiaopeng Fan⁴, Tiejun Huang^{1,2}

¹School of Computer Science, Peking University

²National Key Laboratory for Multimedia Information Processing, Peking University

³School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University

⁴School of Computer Science and Technology, Harbin Institute of Technology

wangyuanlin@stu.pku.edu.cn

{yiyangz, rqxiong, jzhaopku, zhangjian.sz, tjhuang}@pku.edu.cn fxp@hit.edu.cn

8. Model Details

8.1. Convolutional Layer and Residual Block

In the structure of the proposed Spk2SRIImgNet, there are some “convolutional layer”s and “residual block”s. The former indicates a convolution followed by a ReLU activation function, while the latter refers to [1]. The structures are shown in Fig. 10.

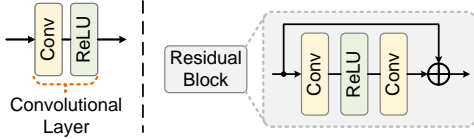


Figure 10. Illustration of convolutional layer and residual block.

8.2. Visualization of Channel Rearrangement.

Steps 2 and 6 in MACF module are two operations that do not involve model parameters. We provide a visualization of step 2 to clarify this operation. The encoder outputs $\{F_i\}_{i=-4}^4$, which are features at different moments. We present 64-channel features of F_0 and F_4 , respectively, indicating that different channel encodes different information, as shown in Fig. 11 (a), (b). The comparison between Fig. 11 (a) and (b) illustrates that features at different moments encode the same information (*e.g.* edges, textures, etc) at the same channel index. For example, F_0^1 and F_4^1 encode the same information at different moments, *i.e.*, $\{F_i^1\}_{i=-4}^4$ are the similar unaligned data. We implement

channel rearrangement to gather these similar unaligned features for subsequent joint processing.

G_j is formed by collecting feature maps from $\{F_i\}_{i=-4}^4$ at the same channel index j . Fig. 11 (c) and (d) show the features collected from channel index 3 and 9, respectively. For example, G_3^0 is collected from F_0^3 ($G_3^0 = F_0^3$); G_3^4 is collected from F_4^3 ($G_3^4 = F_4^3$); G_9^4 is collected from F_4^9 ($G_9^4 = F_4^9$). G_j contains similar temporal unaligned features, which are then processed by steps 3, 4, and 5 to mitigate fluctuations by exploiting their similarity. The visualization helps to understand the motivation behind each step in MACF module.

8.3. Comparison with SpikeSR-Net

The proposed Spk2SRIImgNet leverages the long-term temporal correlation of spikes to enhance the consistency of multi-moment features along the motion trajectory, handling fluctuations and improving SR quality. Specifically, we extend the spatial-domain patch processing to temporal filtering of spike stream and design a novel framework MACF, which incorporates motion alignment, transform domain collaborative filtering, and inverse alignment operations. In contrast, SpikeSR-Net [9] focuses only on the local temporal correlation of spikes to extract features at each moment.

Additionally, our method follows a single-pass architecture, while SpikeSR-Net is an optimization-inspired network with an iterative structure, designed by unfolding the optimization process of SCSR observation model.

8.4. Plural Usage

Since \tilde{G}_j , $\Delta\tilde{G}_j$, $\Delta\tilde{G}_j$, etc., include features from multiple moments, we call them ‘features’ (plural).

*Corresponding author.

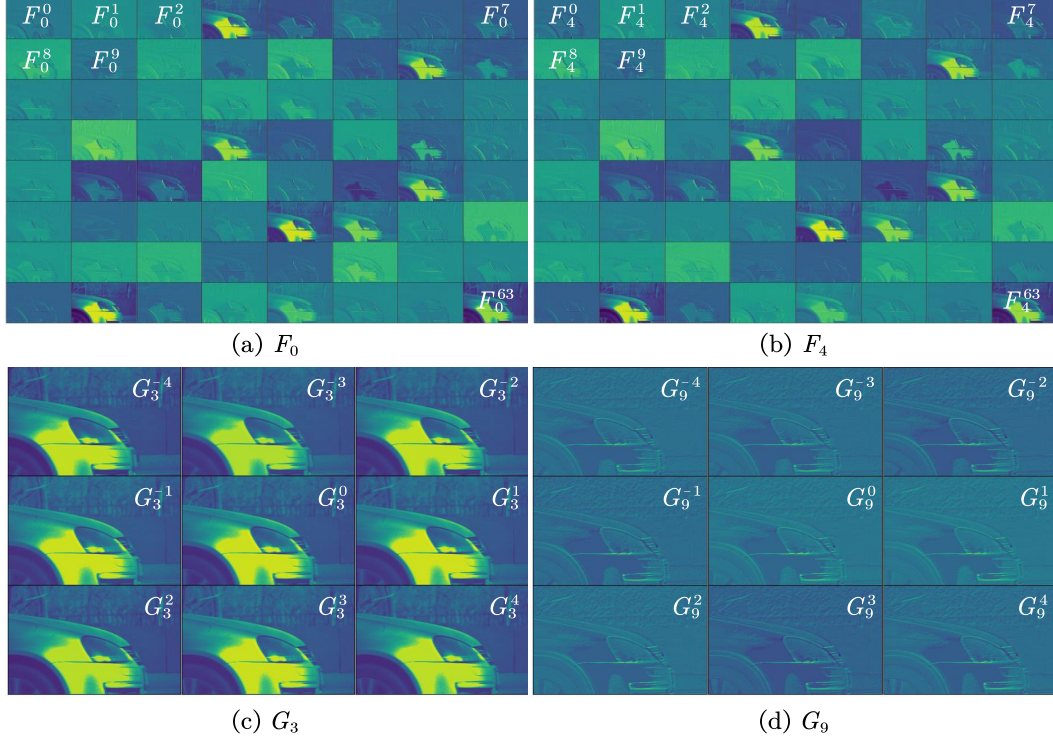


Figure 11. Visualization of channel rearrangement (CR) in MACF module. (a) 64 dimensional feature maps of F_0 ; (b) 64 dimensional feature maps of F_4 ; (c) 9 dimensional feature maps of G_3 ; (d) 9 dimensional feature maps of G_9 .

9. Experimental Details

9.1. Spike Camera Simulator

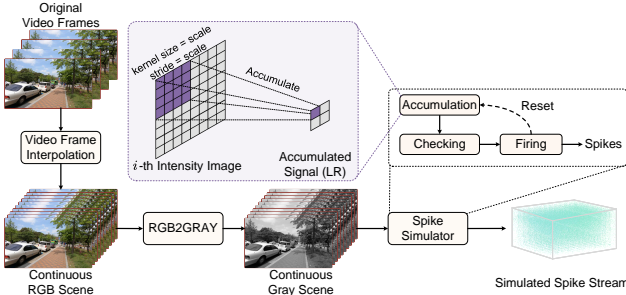


Figure 12. The pipeline of spike camera simulator.

The pipeline of the spike camera simulator is illustrated in Fig. 12. We regard the input video as the dynamic scene to generate the spike stream. Due to the limited frame rate of videos, their temporal information is insufficient to generate spike stream with ultra-high temporal resolution. Therefore, we first use the video frame interpolation method [8] to generate latent intensity frames between original video frames to simulate the continuous scene. Then, we convert the color scenes to gray scenes. The gray video frames with high temporal resolution are fed into the spike

simulator. The spike simulator follows the working mechanism of spike camera to accumulate light intensity from each latent frame and periodically compare the accumulated intensity with a preset threshold for firing spikes.

9.2. Loss Function

we use \mathcal{L}_1 loss to train the model. The loss function is formulated as follows:

$$\mathcal{L}_1 = \|\mathcal{I}(t_0)/\eta - \mathcal{I}^{gt}(t_0)\|_1,$$

where $\mathcal{I}(t_0)$ is the generated high resolution image at time t_0 , η is the photoelectric conversion rate, $\mathcal{I}^{gt}(t_0)$ is the ground truth.

9.3. MACs and Memory Cost

As shown in Tab. 3, we compare the proposed method with other methods on computational complexity (MACs, multiply-accumulate operations) and GPU memory cost. MACs and memory cost are tested on LR input of size $180 \times 360 \times 101$ with $\times 4$ SCSR model.

9.4. More Quantitative Results

Selection of Video Dataset. High-resolution training and testing datasets are essential for super resolution models. First, high-resolution datasets enable the network to learn to

Method	MACs/G	Memory/MB
TFP+DAT	862.0	4234
TFP+BasicVSR++	3636.0	3986
Spk2imgNet+DAT	1434.5	6768
Spk2imgNet+BasicVSR++	8781.4	26720
WGSE+DAT	1120.4	7126
WGSE+BasicVSR++	5732.7	24020
BSF+DAT	1270.3	7474
BSF+BasicVSR++	7263.1	33848
VidarSR	16537.4	6462
SpikeSR-Net	7273.2	5486
Ours	1238.5	7950

Table 3. Comparisons on MACs and GPU memory cost.

Dataset	Scenes	Sample Pairs	Resolution
REDS [3]	30	150	720×1280
Adobe240 [4]	17	170	720×1280
UDM10 [7]	10	10	720×1272
Vid4 [2]	4	4	-
Vimeo-90K-T [6]	7824	7824	256×448

Table 4. Details of the synthesized SCSR evaluation datasets. The ‘sample pairs’ denote the number of samples in SCSR evaluation dataset.

restore finer details, while low-resolution datasets limit the network’s learning ability. Second, HR datasets can effectively evaluate the model performance, while LR datasets struggle to fully reflect the model performance due to the lack of sufficient details.

Thus, we choose REDS [3] and Adobe240 [4] datasets, both with a spatial resolution of 720×1280 . Additionally, their high frame rates (120 fps for REDS and 240 fps for Adobe240) make them well-suited for simulating ultra-high-frame-rate spike data.

Evaluation on Other Datasets. UDM10 [7] is a video super resolution (VSR) test dataset with a high spatial resolution of 720×1272 . Vid4 [2] and Vimeo-90K-T [6] are commonly used test datasets in video restoration tasks. We also generate UDM10-based, Vid4-based, and Vimeo-90K-T-based SCSR datasets for evaluation, as shown in Tab. 6. The details of the synthesized SCSR evaluation datasets are presented in Tab. 4. It is noted that Vimeo-90K-T [6] contains only one HR image per scene. Thus, we generate the corresponding LR spike stream based on the LR image sequence, which differs from the general procedure in Sec. 9.1.

$\times 8$ Quantitative Results. We also conduct experiments on $\times 8$ SR scale. Each sample from $\times 8$ SCSR dataset con-

Case	SIR	MACF	PSNR \uparrow	SSIM \uparrow	Params(M)
(a)			28.85	0.8131	2.50
(b)	✓		28.88	0.8142	2.50
(c)		✓	29.31	0.8304	3.64
(d)	✓	✓	29.37	0.8323	3.64

Table 5. Ablation study on Spk2SRImgNet- ℓ with REDS-based SCSR evaluation dataset on $\times 4$ SR scale.

sists of the spike data of size $90 \times 160 \times 101$ and a ground truth image of size 720×1280 . During training, we randomly crop the spike stream to a spatial size of 48×48 , with other settings the same as in Sec. 6.1. The quantitative results are presented in Tab. 7. Our method demonstrates competitive performance compared with VidarSR [5], while requiring fewer model parameters and achieving faster inference.

9.5. More Ablation Study

Motion Aligned Collaborative Filtering (MACF) Module. In Spk2SRImgNet, the encoder consists of 2 convolutional layers and 3 residual blocks. To further validate the effectiveness of MACF module, we replace the encoder in Spk2SRImgNet with a lightweight encoder ℓ consisting of only 2 convolutional layers, while keeping the other modules unchanged. This results in a lightweight-encoder version of Spk2SRImgNet, denoted as Spk2SRImgNet- ℓ . We conduct ablation study on Spk2SRImgNet- ℓ , as presented in Tab. 5. Comparisons between (a) and (b), (c) and (d) illustrate the advantages of SIR module. Comparisons between (a) and (c), (b) and (d) demonstrate the effectiveness of MACF module, showing gains of 0.46dB and 0.49dB, respectively. Besides, we observe that Spk2SRImgNet- ℓ outperforms two existing SCSR methods: VidarSR [5] and SpikeSR-Net [9].

Other Operations. We perform ablation study on some operations in Spk2SRImgNet, as presented in Tab. 8. Case (5) is the proposed Spk2SRImgNet. In case (1), the alignment operation based on deformable convolution in motion aligned super resolved reconstruction (MASR) module is removed from Spk2SRImgNet to demonstrate the effectiveness of this design. In case (2), the aggregation operation (i.e., Eq. (11)) in step 5 of MACF is removed. Comparisons between cases (2) and (5) illustrate the effectiveness of the aggregation operation. In case (3), the noise estimation (NE) submodule in step 4 of MACF is removed from Spk2SRImgNet to verify its contribution. In case (4), removing CR operation (steps 2 and 6) from MACF degrades the performance, thereby demonstrating its role.

Scale	Method	UDM10-based Dataset			Vid4-based Dataset			Vimeo-90K-T-based Dataset			Params(M)	Runtime(s)
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		
×4	VidarSR	33.13	0.9038	0.2011	23.95	0.7444	0.3275	28.50	0.8714	0.2360	12.79	0.11
	SpikeSR-Net	33.93	0.9173	0.1838	24.62	0.7833	0.3052	28.59	0.8765	0.2338	3.34	0.29
	Ours	34.20	0.9202	0.1743	24.76	0.7893	0.2982	28.52	0.8723	0.2322	3.86	0.04

Table 6. Quantitative results on UDM10-based, Vid4-based and Vimeo-90K-T-based SCSR dataset on ×4 super resolution scale. **Red** and **blue** indicate the best and the second-best performance, respectively. The runtime is tested on LR input of size $64 \times 112 \times 101$ (×4 SR scale).

Scale	Method	REDS-based Dataset			Adobe240-based Dataset			Params(M)	Runtime(s)
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓		
×8	VidarSR	25.72	0.6855	0.4261	26.36	0.7597	0.3205	20.33	0.69
	SpikeSR-Net	25.67	0.6801	0.4397	26.20	0.7503	0.3414	3.59	0.90
	Ours	25.85	0.6883	0.4318	26.39	0.7576	0.3339	4.01	0.07

Table 7. Quantitative results on REDS-based and Adobe240-based SCSR dataset on ×8 super resolution scale. **Red** and **blue** indicate the best and the second-best performance, respectively. The runtime is tested on LR input of size $90 \times 160 \times 101$ (×8 SR scale).

Case	Setting Description	PSNR ↑	SSIM ↑
(1)	Removing alignment in MASR	29.21	0.8267
(2)	Removing aggregation in step 5 of MACF	29.37	0.8326
(3)	Removing NE in step 4 of MACF	29.46	0.8353
(4)	Removing steps 2 and 6 in MACF	29.43	0.8343
(5)	The final model	29.50	0.8369

Table 8. Ablation study on some operations in Spk2SRImgNet with REDS-based SCSR evaluation dataset on ×4 SR scale.

K	T_w	PSNR ↑	SSIM ↑	Params(M)	Runtime(s)
1	21	27.91	0.7809	2.07	0.10
3	41	28.83	0.8126	2.45	0.13
5	61	29.17	0.8243	3.12	0.16
7	81	29.35	0.8311	3.49	0.19
9	101	29.50	0.8369	3.86	0.21
11	121	29.57	0.8394	4.24	0.25

Table 9. Ablation study on the number of input spike frames with REDS-based SCSR evaluation dataset on ×4 SR scale.

Number of Input Spike Frames. We set the number of partitioned short-term spike blocks K to $\{1, 3, 5, 7, 9, 11\}$, i.e., the number of input spike frames T_w is $\{21, 41, 61, 81, 101, 121\}$. The corresponding results are presented in Tab. 9. In this paper, we choose $K = 9$, i.e., $T_w = 101$ (bold in Tab. 9).

9.6. More Visual Results

To enhance understanding of spike camera super resolution task, we provide a *video* showcasing the continuous SR re-

construction results on real-captured spike data. As there is no ground truth for these real-world spike data, we use two basic reconstruction methods: TFI [10] and TFP [10] to provide image reconstruction results for reference. TFI is the spike interval based reconstruction (SIR) method, which is introduced in Sec. 3. TFP accumulates photons in a virtual exposure window, similar to the exposure window based imaging model of conventional camera. We implement TFP with a temporal window of 30. The TFP results can serve as references for raw grayscale images captured within a very short exposure window.

Besides, we provide more visual comparison results on real-world spike data and synthesized SCSR data to further demonstrate the superior performance of the proposed Spk2SRImgNet, as shown in Figs. 13 to 18. Compared with other methods, the proposed model generates HR images with better details and visual quality. Please enlarge the figure for better comparison.

References

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1
- [2] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):346–360, 2014. 3
- [3] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

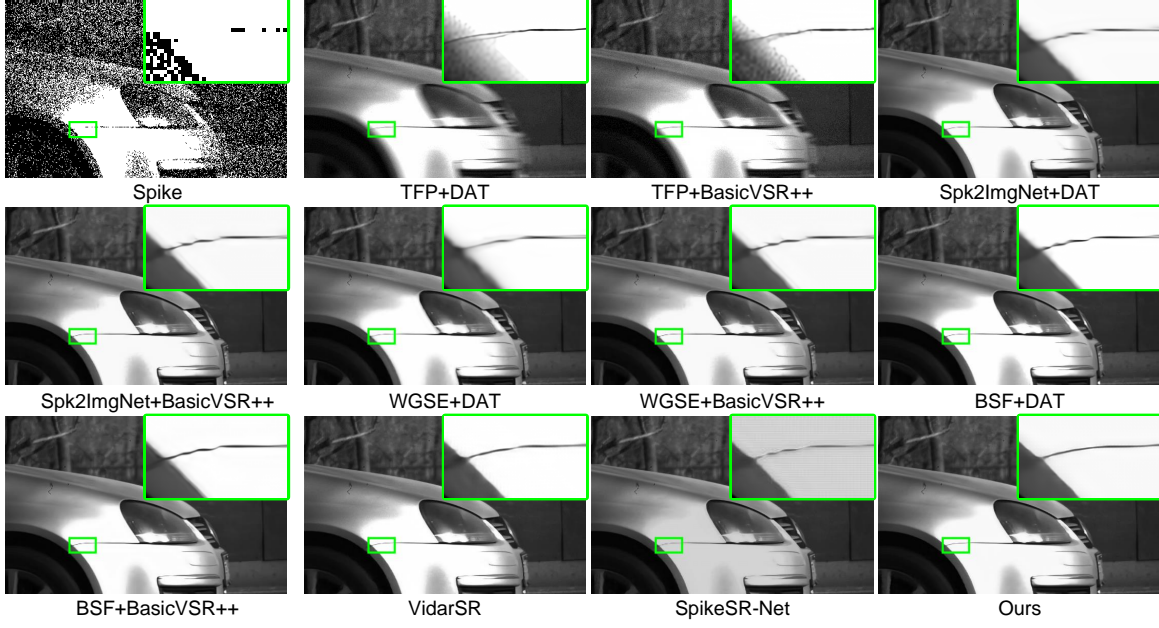


Figure 13. Visual comparison ($\times 4$) on real-world spike data. The spike stream records a running car. Please enlarge the figure for better comparison.

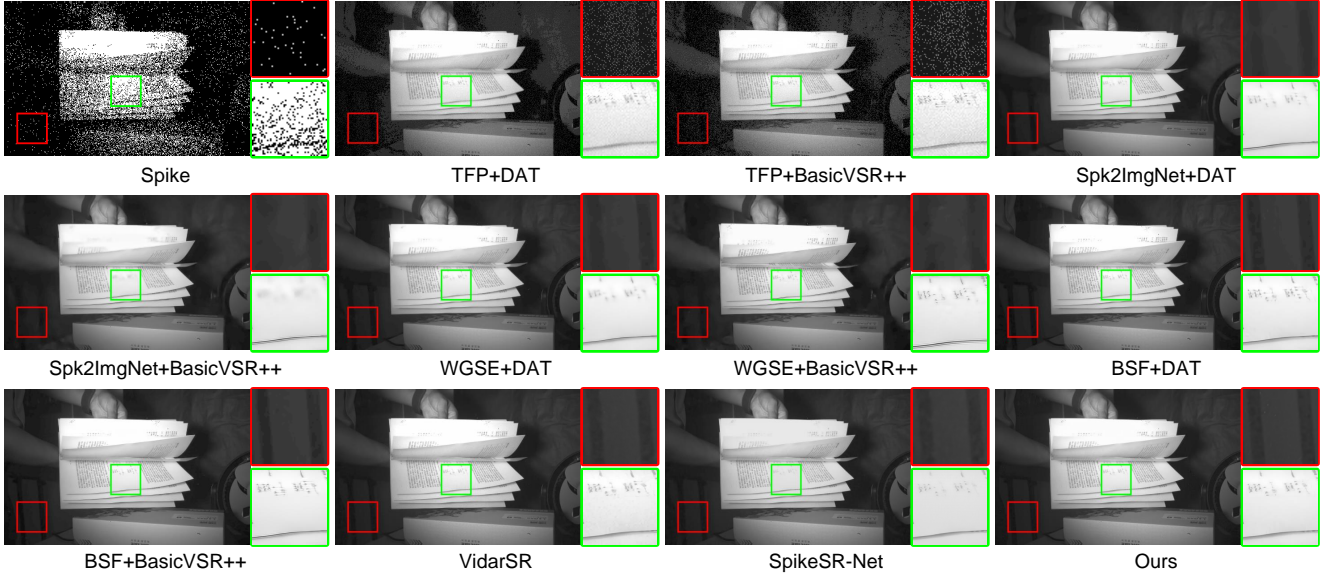


Figure 14. Visual comparison ($\times 4$) on real-world spike data. The spike stream records a scene involving flipping through pages. Please enlarge the figure for better comparison.

- [4] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1279–1288, 2017. 3
- [5] Xijie Xiang, Lin Zhu, Jianing Li, Yixuan Wang, Tiejun Huang, and Yonghong Tian. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE*

Transactions on Circuits and Systems for Video Technology, 33(1):16–29, 2023. 3

- [6] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 3
- [7] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. Progressive fusion video super-resolution network



Figure 15. Visual comparison ($\times 4$) on the REDS-based spike data. The values below each image represent “PSNR / SSIM” metrics. Please enlarge the figure for better comparison.

- via exploiting non-local spatio-temporal correlations. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3106–3115, 2019. 3
- [8] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5682–5692, 2023. 2
- [9] Jing Zhao, Ruiqin Xiong, Jian Zhang, Rui Zhao, Hangfan Liu, and Tiejun Huang. Learning to super-resolve dynamic scenes for neuromorphic spike camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3579–3587, 2023. 1, 3
- [10] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture recon-

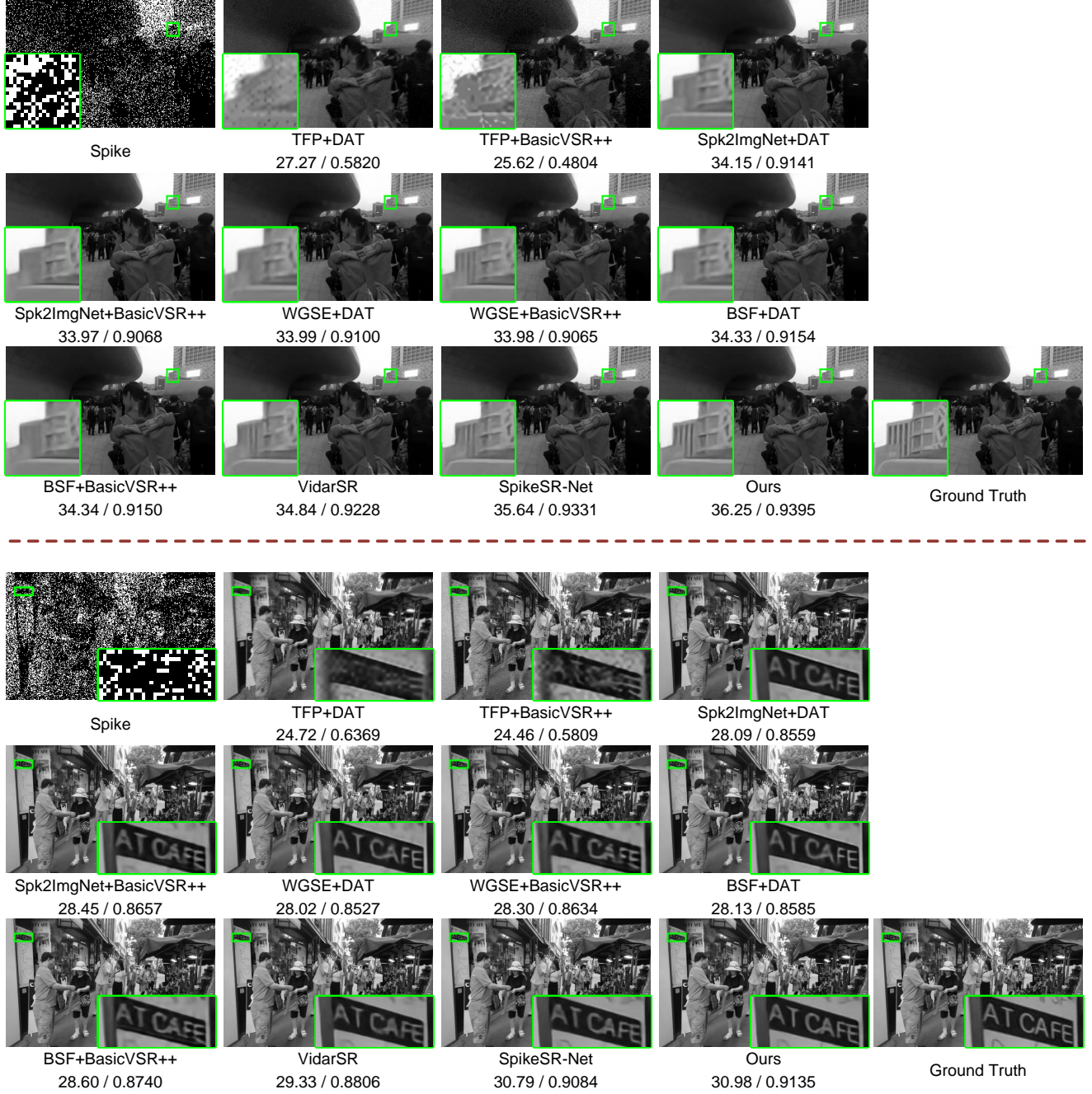


Figure 16. Visual comparison ($\times 4$) on the REDS-based spike data. The values below each image represent “PSNR / SSIM” metrics. Please enlarge the figure for better comparison.

struction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1432–1437. IEEE, 2019.



Figure 17. Visual comparison ($\times 4$) on the Adobe240-based spike data. The values below each image represent “PSNR / SSIM” metrics. Please enlarge the figure for better comparison.

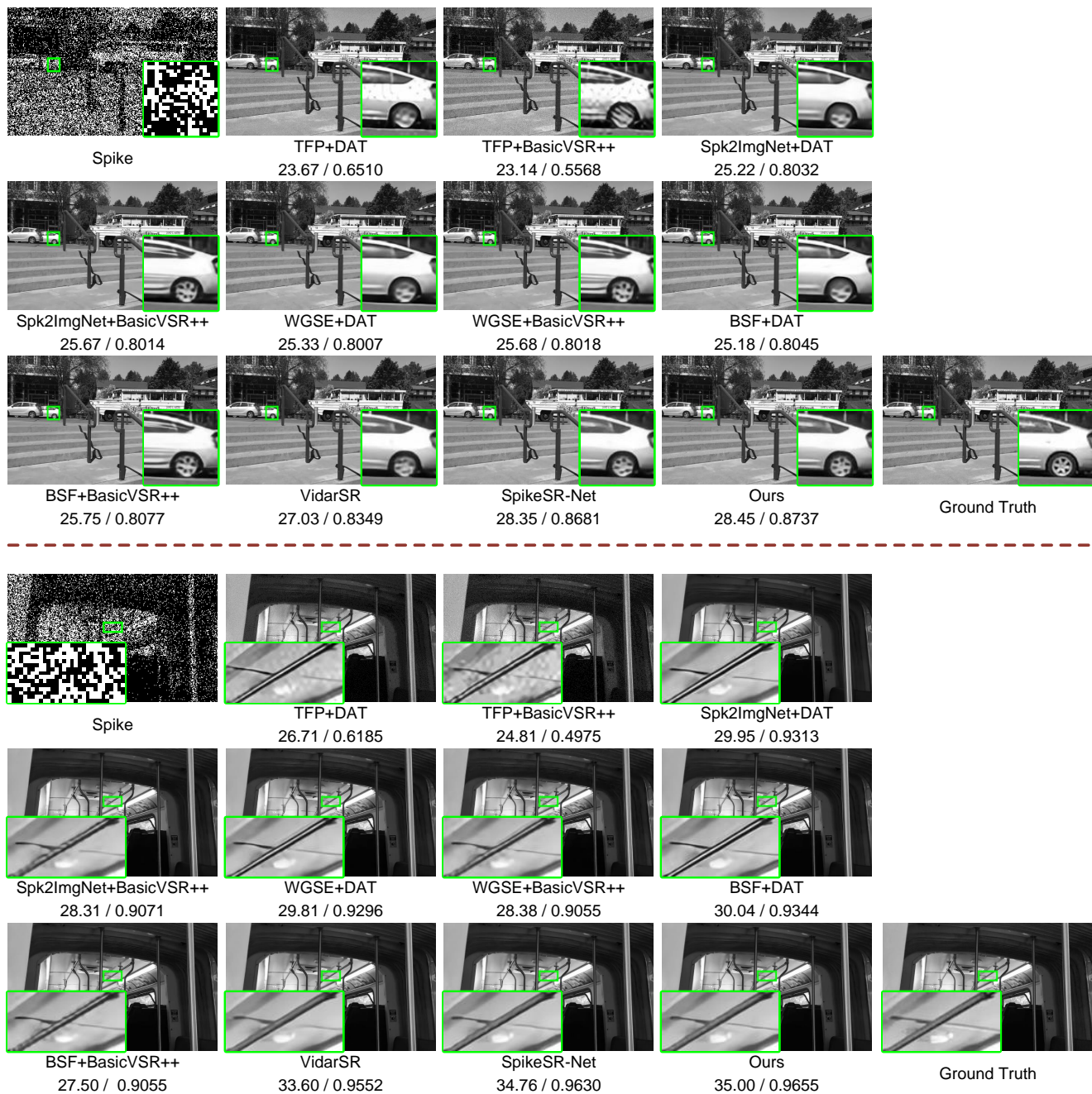


Figure 18. Visual comparison ($\times 4$) on the Adobe240-based spike data. The values below each image represent “PSNR / SSIM” metrics. Please enlarge the figure for better comparison.